The Attention Hypothesis

A Theory of the Fermi Paradox as Internal Filter

A.H.

2026

A Companion to Two Kinds of Conditioning

Abstract. This paper proposes a theory of the Fermi paradox as an internal filter rather than an external one. The standard candidates for a filter ahead of technological civilizations are catastrophic events arriving from outside: asteroid impacts, nuclear exchange, engineered pathogens, ecological collapse. These explanations share a common structural weakness: to account for the near-total absence of detectable civilizations across a galaxy of hundreds of billions of star systems, any external filter would have to operate with a universality that no known physical mechanism possesses. This paper proposes a different candidate. The filter is internal to the structure of the cognitive instrument that makes civilization possible in the first place. Specifically, it proposes that the same biological mechanism that produces technological civilization contains, as a structural feature of how it operates, a compounding dynamic that progressively eliminates the quality of attention required to use technology wisely. This dynamic, once running beyond a certain developmental threshold, is structurally resistant to correction by the very apparatus that would be needed to correct it. The theory does not predict catastrophe. It predicts a particular kind of quiet failure, the most sophisticated kind, the kind that presents at every stage as progress. Crucially, the link from this internal degradation to cosmic silence is not assumed but argued: detectability is expensive, and the same insensitivity that degrades a civilization is what consumes the surplus that detectability requires. The theory carries a corollary: the filter is, in principle, passable, and the conditions for passing it are specifiable. It also carries a specific and unresolved exposure, examined directly rather than hidden: persistent automation that no longer needs the healthy substrate. This paper develops the theory in full, states its premises explicitly, identifies what would disconfirm it, and draws out its implications for the current moment.

I. The Problem

In 1950, physicist Enrico Fermi asked a question at lunch that has not been satisfactorily answered since. Given the age of the universe, approximately 13.8 billion years, given the abundance of star systems with planets in conditions apparently favorable for life, given the

billions of years available for technological civilizations to develop and expand and make themselves detectable across interstellar distances, why is there no evidence of any civilization other than our own? The silence is not what the numbers predict. It is, statistically, strange.

The Drake equation, formulated by astronomer Frank Drake in 1961, attempted to estimate the number of active communicating civilizations in the Milky Way galaxy by multiplying a series of factors: the rate of star formation, the fraction of stars with planets, the fraction of those planets capable of supporting life, the fraction on which life actually develops, the fraction on which intelligent life develops, the fraction that develop detectable technology, and the average length of time such civilizations remain detectable. Estimates of the final product range from less than one to millions, depending on assumptions. What the silence suggests is that the product is very small, perhaps approaching zero for detectable civilizations in our galaxy at this moment. Something is reducing the number drastically. The question is what.

Robin Hanson formalized this problem in 1998 with the concept of the Great Filter: something filters civilizations out before they become detectable. The filter could be behind us, in the development of life, of complex multicellular organisms, of intelligence itself, in which case we are the rare survivors of an extraordinarily improbable sequence and may be effectively alone. Or the filter is ahead of us, in something that civilizations encounter at a stage of development we are approaching or have reached. This second possibility has focused attention on catastrophic external events: asteroid impacts, gamma-ray bursts, nuclear exchange, resource depletion, engineered pathogens, climate collapse. These are real risks. They receive serious attention in the literature. But they share a structural problem as candidates for the filter.

For any external event to filter out civilizations with the consistency required to account for the silence, it would have to occur with extraordinary regularity across incomprehensibly diverse conditions. There are hundreds of billions of star systems in the Milky Way alone, and the observable universe contains an estimated two trillion galaxies. The external filter would have to operate as a near-universal law, eliminating technological civilizations regardless of their location, their physical environment, their specific history, their particular resources, their individual circumstances. No known physical mechanism has that property. Asteroid impacts are probabilistic and depend on orbital dynamics specific to each planetary system. Nuclear exchange requires a specific political and technological history. Engineered pathogens require a specific biological and technological convergence. Each of these is contingent in ways that preclude the universality the silence demands.

This is the gap the theory in this paper attempts to fill. If the filter is internal rather than external, if it follows from the structure of the cognitive instrument itself rather than from contingent

physical circumstances, then it would operate across diverse conditions with the near-universality that the silence requires, because the instrument, not the environment, is the common factor.

II. The Instrument and Its Structural Failure Mode

The theory rests on a distinction between two kinds of cognitive conditioning, developed at length in the companion investigation *Two Kinds of Conditioning*. A compressed account is given here sufficient to make the theoretical argument, but the full development, including the evidentiary basis in comparative cognition, neuroscience, and behavioral data, is in that work.

The first kind is biological conditioning. This is the cognitive instrument doing what it was built to do: accumulating knowledge through experience, building models of the physical and social world, solving the problems of survival at individual and collective scale. Language, tool use, planning, anticipation of consequences, navigation of complex social environments, the accumulation and transmission of knowledge across generations: all of this is biological conditioning operating in its domain. It is necessary and good. It is what makes civilization possible.

The second kind is psychological conditioning. This is the same instrument turned on things outside its domain. Using the problem-solving apparatus to solve the problem of longing. Using the categorization system to categorize love. Using the optimization function to optimize for meaning. Using the framework-building capacity to build a framework for what it feels like to be alive. The instrument produces outputs. The outputs have no relationship to the thing they were aimed at, because that thing was never in the domain of the instrument. But the outputs look like progress. They feel like progress. They are the product of a powerful and well-functioning apparatus, and the apparatus does not have a mechanism for detecting when it has exceeded its domain, because the detection would require a capacity the apparatus itself cannot supply.

This misapplication produces four compounding mechanisms. The first is the narrowing of perception through accumulated expertise and category: the expert becomes progressively less able to encounter genuinely new information in their domain because the architecture of their expertise has become the architecture of their perception, and what falls outside the category system does not register as unknown but simply does not register. The second is the self-exemption reflex: when a principle is recognized as applying to the self in an uncomfortable way, the reflex generates, before deliberation, a reason why the self is the exception, using the understanding of the principle as the material for the most convincing possible exemption from it. The third is the pacification of urgency: the system absorbs opposition by giving it just enough room to discharge that it stops generating real pressure, so that the most capable people become

the most effectively pacified, because they have the most sophisticated apparatus for constructing the internal condition that makes the promise of solution feel real enough to dissolve the urgency. The fourth is the progressive elimination of the quality of attention required to interrupt any of the other three: the narrowing, the reflex, and the pacification together create conditions in which the quality of presence that would be needed to see the mechanisms operating is the quality of presence that none of the mechanisms will permit.

These four mechanisms compound each other. The narrowing makes the reflex more effective. The reflex makes the pacification more complete. The pacification reinforces the narrowing. The whole structure is self-reinforcing and becomes more stable, not less, as it develops. It is not a pathology. It is the natural consequence of a very powerful instrument encountering things outside its domain and having no internal check on the exceeding.

The developmental threshold at which this becomes civilizationally significant is the point at which technological capacity outpaces the natural friction that previously slowed the misapplication. In pre-technological conditions, the consequences of misapplying the instrument to things outside its domain were local and relatively slow. Mistakes had direct personal consequences that kept the self-exemption reflex partially in check. The pace of change was slow enough that the quality of attention could occasionally reassert itself through the ordinary experience of being wrong in ways you could not avoid noticing. As technological capacity increases, the separation between decision and consequence widens, the pace of change accelerates, and the conditions under which the mechanisms could be interrupted become progressively rarer. The instrument produces more power more quickly, and the misapplication of that power becomes harder to detect and harder to correct, by the same mechanisms that are producing the power.

III. The Universality Premise and Its Status

The theory requires a premise that must be stated honestly and examined carefully, because the argument stands or falls on it. The premise is this: the four mechanisms described above are not peculiar features of human cognition, products of our specific evolutionary history or our particular social arrangements or our contingent cultural development. They are structural features of any sufficiently powerful problem-solving instrument encountering things outside its domain.

This premise is not established. It cannot be established from the evidence available to us, which consists of exactly one example of a technological civilization. We cannot examine other intelligences to determine whether they face the same dynamic. We cannot run the experiment.

What we can do is examine the premise on its structural merits and ask whether there are strong reasons to think it would or would not generalize.

The argument for generalization is this. The mechanisms do not arise from specifically human features of cognition. They arise from the relationship between any powerful problem-solving instrument and the existence of things outside its domain. The narrowing of perception through accumulated category is a structural consequence of the instrument working well: the better it works, the more its categories dominate its perception, and the more it fails to register what falls outside them. This is not a bug. It is a feature of efficient categorization, and any efficient categorization system would face it. The self-exemption reflex arises from the relationship between self-interest and reasoning: any system that reasons about its own interests will generate post-hoc justifications for conclusions it is already disposed toward. This is not a specifically human cognitive bias. It is a structural consequence of having interests and the capacity to reason about them. The pacification of urgency is a consequence of the gap between perceiving a problem and bearing its consequences: any system in which decision and consequence can be separated will face this. The elimination of the quality of attention is a consequence of the other three operating together, and would follow for any system in which they operated.

The argument against generalization is that we do not know what other forms intelligence might take, and radically different architectures might not face the same dynamics. An intelligence that was genuinely collective rather than individual, with no self-exemption reflex because there is no self to exempt, might navigate differently. An intelligence that did not develop through the evolutionary pressures that shaped the human relationship between reasoning and self-interest might not produce the same distortions. A form of cognition sufficiently alien to our own might find the territory outside the instrument’s domain through entirely different means.

Both arguments are honest and neither is conclusive. The theory holds the premise as a serious possibility rather than an established fact, and acknowledges that this is where the theory is most exposed at the level of the mechanism. What can be said with confidence is this: if the premise is false, if there are forms of technological intelligence that do not face this dynamic, then the filter is not universal, and the silence has some other explanation. If the premise is true, the theory accounts for the silence more elegantly than any external candidate, because it is the only candidate that does not require a contingent physical mechanism to operate universally.

There is, however, a second universality burden that the generalization of the mechanisms does not by itself discharge, and it is worth naming here before it is developed. Even granting that the four mechanisms arise in any sufficiently powerful instrument, the theory must still explain why the degradation they produce reliably issues in silence rather than in some loud and durable

artifact that outlives the degradation. That is a question not about whether the mechanisms generalize but about whether they close every route to detectability. It is taken up directly in Section V, and it is, in the end, the place where the theory is most exposed of all.

IV. The Shape of the Failure

The filter, as this theory describes it, would produce a characteristic signature that distinguishes it from external catastrophe candidates. Understanding this signature matters both for evaluating the theory and for recognizing its operation in real time.

The failure would not be sudden. It would be progressive and self-concealing. At every stage, the civilization would appear to be advancing. Technological capability would continue to increase. Material conditions for many would continue to improve. The complexity and sophistication of the civilization's outputs, its art, its institutions, its scientific understanding, its infrastructure, would continue to grow. The failure would be occurring at the level of the quality of attention available for the navigation of that complexity, not at the level of the complexity itself. The instrument would become more powerful while the capacity to use it wisely was being progressively eliminated by the mechanisms the instrument's own power was producing.

The failure would be invisible from inside it to those most thoroughly shaped by it. The people most embedded in the mechanisms would experience themselves as the most knowledgeable, the most sophisticated, the most capable of understanding the situation. Their narrowed perception would present as expertise. Their self-exemption would present as nuanced understanding of why their particular case is different. Their pacified urgency would present as mature acceptance of complexity. Their eliminated attention would present as efficiency. The failure would look, from inside, like flourishing.

The failure would accelerate rather than plateau. As the mechanisms compound each other, and as the technological capacity of the civilization increases the speed and scale at which the misapplication operates, the rate of narrowing would increase. The gap between the instrument's power and the quality of attention available to guide it would widen faster with each generation. The civilization would reach a point at which its technological capacity was extraordinary and its capacity to navigate that technology wisely was nearly exhausted, and the gap between the two would be the largest it had ever been, and widening.

The failure would produce no visible catastrophe to detect from outside. The civilization would not end in an explosion or a plague or a collision. It would end in a progressive loss of the capacity to sustain the kind of patient, distributed, collective attention that large and slow

undertakings require. Not a bang. Not even a whimper. A gradual, sophisticated, self-justifying narrowing, until the thing that could have heard what it needed to hear was no longer present in sufficient measure to make the difference.

The silence of the universe, on this theory, is not the silence of disaster. It is the silence of a particular kind of ending that does not announce itself and leaves no dramatic wreckage: the silence of civilizations that became very good at solving problems and then tried to solve everything, including the things that were never problems, and in doing so progressively eliminated the capacity to hear what was never a problem to begin with. But the claim that this kind of ending produces silence, rather than a quiet civilization surrounded by the loud and lasting evidence of what it once built, is not self-evident. It is the claim the next section has to earn.

V. The Cost of Being Heard

The argument to this point carries a hidden assumption, and the theory is stronger for dragging it into the open. The assumption is that a degrading civilization would still be detectable, so that its degradation would register, from outside, as silence. If detectability were free, this would be unobjectionable. It is not free, and the easy version of the claim should be abandoned first. Detectability is not leakage. A civilization's incidental emissions do not announce it across the galaxy; our own radio and television signals disperse and fall into the noise floor within a few light-years of Earth, and no plausible receiver reconstructs them at interstellar range. Nobody is listening to our broadcasts, because there is, in any meaningful sense, nothing left to hear. Whatever silence the universe presents, it is not the absence of overheard chatter.

Actual technosignatures, the kind a search could find, are deliberate and expensive. They are sustained beacons aimed and powered across centuries; they are megastructures large enough to alter a star's output, of the kind Dyson described in 1960 and Kardashev built into his scale of civilizational energy use in 1964; they are the physical evidence of expansion across a system or beyond it. Every one of these is not an emission but a project. And every such project requires, simultaneously, three things: surplus energy far beyond what subsistence and ordinary complexity consume; deep cooperation sustained across institutions and across generations, because none of these undertakings can be completed within the attention span of the people who begin them; and a stable biosphere underneath the whole effort, because the surplus and the cooperation both rest on it. To be heard is not to be loud. It is to have built, and kept building, something that only a civilization with energy to spare, patience to spend, and a living foundation beneath it could build at all.

This is where the internal filter does its work without needing a separate step, and it is where the abstract language of the previous section, the loss of a capacity for collective attention, becomes a concrete causal chain. The degradation the theory describes is precisely the loss of sensitivity to slow, distributed, unglamorous costs. Consider a rainforest. Its destruction is not a decision anyone makes; it is the sum of a vast number of locally reasonable choices whose costs are diffuse, deferred, and borne by no one at the table. The insensitivity that permits this is not a mood or a failure of values. It is the inability of the decision-making structure to register a cost that does not arrive as a discrete, legible, present-tense problem of the kind the instrument is built to solve. And that same insensitivity, operating everywhere at once, is what consumes the surplus. The externalities are not registered, so they are not priced, so they compound, until the energy and cooperation and biospheric stability that a cosmic-scale project would have run on are instead spent, permanently and with rising urgency, servicing the accumulating wreckage. The attention collapse is the coordination collapse is the energy collapse. They are not three failures but one, seen from three distances. A civilization on this path does not go quiet by exploding. It goes quiet by becoming permanently too busy with self-inflicted triage to ever build the loud thing. The silence is not the mark of its death. It is the mark of its having been, from a certain point onward, fully occupied.

This is a better account than the previous section's, and it should be allowed to do its work before its cost is paid. But it incurs a cost, and intellectual honesty requires paying it here in full rather than deferring it. Making detectability expensive closes the easy objection and opens a harder one: automation that no longer needs the healthy substrate. The theory itself insists that the failure presents at every stage as progress, which means there is a long window in which a civilization's raw capability is extraordinary while its wisdom is already draining away. The loud thing need only be launched once, inside that window. A self-replicating probe of the kind von Neumann's work on self-reproducing automata made conceivable, and Tipler later pressed into the Fermi debate, or even dumb but persistent machine infrastructure of the kind Bracewell imagined, does not need ongoing cooperation, sustained attention, or a living biosphere once it is under way. It expands on its own. So for the silence to hold, it is not enough that civilizations lose the capacity to keep building the loud thing. It must be that none of them, anywhere in hundreds of billions of star systems, ever seeds a single self-sustaining thing during its capable-but-degrading phase. The filter has to close that one escape hatch every time, everywhere. One exception, in the entire history of the galaxy, is enough to break it.

There is a reply, and the theory should make it while being candid that it is a reply and not yet a proof. The reply is that the same insensitivity that cuts down the forest also cannot sustain the institutional patience that a century-scale launch demands. Attention collapse, on this account,

does not strike capability uniformly; it strikes the long-horizon, deferred-payoff, distributed-cost undertakings first, because those are exactly the slow and unglamorous commitments the degraded decision structure can no longer register or fund. The disposition required to conceive, resource, and see through a self-sustaining interstellar project across generations is, on this reading, the very first thing the filter consumes, and it is gone before raw capability peaks, not after. Sagan and Newman argued in 1983 that a civilization wise and old enough to build self-replicating probes would be wise enough not to fill the galaxy with them; the internal filter turns the same intuition to a darker purpose, suggesting that the patience and coherence such a project demands are precisely what the degradation removes earliest. This is plausible. It is consistent with everything else the theory claims. But it is a claim that has to be argued, and the argument is not yet decisive, because it asks us to believe something specific about the order in which capacities fail, namely that the patient ones fail first, and that ordering is asserted here more than it is demonstrated. The energy argument makes the quiet failure quieter than a naive reading would credit, and that is a real gain. What it does not yet do is make the silence universal, and universality is the entire job a Fermi filter has to do.

VI. Evidence Within Our Own Civilization

The theory is not merely speculative. It makes claims about a mechanism that should be observable in a civilization currently undergoing the process the theory describes. We are that civilization. The evidence for the mechanism operating in real time is substantial and measurable.

The narrowing of perception is measurable in the compression of inquiry. From 2000 to 2018, the average length of an internet search query gradually increased, suggesting an expanding aperture of inquiry. Since approximately 2022, the trend has reversed: the median query has shortened, the majority of searches now terminate at an AI-generated summary rather than a primary source, and average time spent engaging with informational content has fallen to under a minute. The perceptual aperture is contracting in real time at population scale.

The self-exemption reflex is measurable in the gap between stated values and behavior. In large-scale sentiment analysis of 2025 social media, users overwhelmingly endorsed principles of ethical consumption and digital wellness. In the same period, ultra-fast fashion platforms and centralized logistics operations hit record traffic and conversion rates. Ethical brand sentiment increased by 40% while behavior moved in the opposite direction. When surveyed, users provided sincere explanations for the discrepancy. The internal condition was generated before the question was asked.

The pacification of urgency is measurable in the substitution of representation for action. Low-bandwidth virtue, the digital representation of moral engagement, surged in the period that high-bandwidth action, sustained reading, donation, physical presence, declined. The representation of having done something has become sufficient to discharge the urgency that would otherwise produce the doing. The urgency is pacified not by being addressed but by being given a form that feels like addressing it.

The elimination of the quality of attention is measurable in the progressive intolerance for friction. More than half of users abandon a webpage that takes more than three seconds to load. This is not a preference for efficiency. It is an intolerance for the duration of discomfort before the reflex fires and the attention moves elsewhere. What is being measured is the progressive elimination of the capacity to remain present to something that does not immediately resolve into a recognizable pattern.

This same insensitivity is visible at the civilizational scale the previous section described, in the systematic mispricing of slow and distributed costs: the externalities that are not registered because they do not arrive as legible present-tense problems, and that compound precisely because the decision structures cannot attend to what does not resolve into a recognizable pattern. The mechanism measurable in the three-second webpage and the mechanism that consumes a civilization’s long-horizon surplus are the same mechanism operating at different scales.

Artificial intelligence, as currently deployed, is measurable as an amplifier of all four mechanisms simultaneously. It is trained on the outputs of the psychological conditioning and produces outputs statistically consistent with that conditioning. It provides the feeling of understanding without requiring the quality of attention that genuine understanding requires. It accelerates the narrowing, provides more sophisticated material for the reflex, offers more comprehensive pacification, and further eliminates the conditions in which the quality of attention could arise. The four-year window from 2022 to 2026 in which the measurable indicators shifted most sharply coincides exactly with the period of mass AI deployment. The timeline of the mechanisms has compressed. What took millennia now takes less.

None of this evidence confirms the theory as applied to the Fermi paradox. It confirms the mechanisms are real and operating. The extension to the paradox requires the universality premise, which remains, as stated, a serious possibility rather than an established fact. But a theory whose proposed mechanism is not only coherent but observably active in the civilization proposing it is on stronger ground than a theory whose mechanism is purely hypothetical.

VII. The Corollary: The Filter Is Passable

Every serious theory of the Fermi paradox as a filter ahead of us carries an implicit or explicit claim about whether the filter is passable. External catastrophe theories generally imply that the filter is passable only by luck: you survive if the asteroid misses, if the pandemic does not emerge, if the weapons are not used. The filter as external event is essentially a lottery.

The internal filter theory implies something different and more specific. The filter is passable, and the conditions for passing it are not a matter of luck. They are a matter of a specific cognitive capacity being developed and maintained at sufficient scale, before the mechanisms eliminate the conditions for its development.

The capacity is the discernment between biological and psychological conditioning: the recognition, in each moment, of whether the instrument is in its domain or exceeding it. This is not a spiritual achievement. It is not the product of a tradition or a practice or an institution. It is not something that can be optimized or scaled or manufactured. It is a quality of attention, available in any ordinary moment to any person willing to stop filling the silence before the silence has delivered what it contains. It was described with precision by Simone Weil as the rarest and purest form of generosity: not concentration, which is the instrument focused on its object, but the receptive, still, non-grasping presence that receives what is actually there rather than confirming what was already decided is there.

The theory's corollary is therefore this: the first civilization to develop this capacity at sufficient scale, before the mechanisms eliminate the conditions for it, would be the first to pass through the filter. Not by solving the problem, because the psychological conditioning is not a problem in the domain of the instrument and cannot be solved by the instrument. By recognizing and interrupting the misapplication, at the moment of misapplication, using the one capacity that stands outside the instrument's operation.

What sufficient scale means is not precisely specifiable. It does not mean universal. The mechanisms do not require unanimity to run. They require only that the people and institutions making the decisions that determine the civilization's trajectory are sufficiently shaped by the mechanisms that the quality of attention required to navigate those decisions is not available in the relevant places at the relevant times. The threshold for passing the filter is not that everyone develops the capacity. It is that enough people in enough positions of consequential decision develop it that the mechanisms no longer run unchecked at the civilizational level.

It is worth noting what Section V implies for this one. If the degradation consumes long-horizon, distributed-cost undertakings first, then the capacity this section describes is not merely a private

corrective; it is the precondition for exactly the patient, surplus-generating, biosphere-sustaining coordination that being heard requires. The capacity that would pass the filter and the capacity that would let a civilization build something lasting are, on this account, the same capacity. To pass through is not separate from being able to build the loud thing. It is what would make the loud thing possible at all.

The theory makes no prediction about whether this civilization will pass the filter. It describes the mechanism and notes that the mechanism is real, that it is running, and that it is accelerating. Whether it runs to completion depends on factors that are not determinable in advance. The investigation is not claiming the outcome is determined. It is claiming that the outcome is not a matter of luck, that it is a matter of a specific capacity, and that the capacity is specifiable, and that the conditions for its development are currently being consumed at an accelerating rate.

VIII. What Would Disconfirm the Theory

A theory that cannot be disconfirmed is not a theory. The internal filter hypothesis makes claims that could, in principle, be shown to be wrong. Stating them explicitly is part of what it means to take the theory seriously.

The theory would be disconfirmed, or significantly weakened, by any of the following.

First: the detection of a technological civilization elsewhere. If we detect a civilization, even one that appears to have navigated past our current developmental stage, the filter question changes entirely. The silence that the theory is attempting to explain would no longer be silence. The theory would need to account for why that civilization survived, which might be possible within the framework, but the evidentiary situation would be fundamentally different.

Second: a compelling account of the silence that does not require a filter at all. If the silence turns out to be fully explained by the limitations of our search, by the size of the galaxy relative to the signals we have examined and the time we have been looking, then the paradox dissolves and neither this theory nor any other filter theory is required. This is a genuine possibility that the theory acknowledges. It is not dismissed; it is simply not sufficient, in the current state of evidence, to make the filter question go away.

Third: a demonstration that the four mechanisms do not generalize, that the structural argument for their universality fails, that there are forms of technological intelligence that do not face the same dynamic. This is one of the two premises the theory is most exposed on, and honest engagement with the theory requires holding this exposure seriously. If it can be shown that the

mechanisms are artifacts of specifically human cognition rather than structural features of any powerful problem-solving instrument, the universality the theory requires is undermined.

Fourth: a demonstration that persistent automation provides a reliable route to durable detectability that does not depend on the healthy substrate. This is the other premise the theory is most exposed on, and it is, on balance, the sharper of the two. The theory's account of silence requires that a civilization in its capable-but-degrading window does not, and effectively cannot, seed self-sustaining infrastructure, self-replicating probes or dumb but persistent expansion machinery, that would outlive the attention required to launch it and continue producing a detectable signature without further cooperation. The theory's defense is that the disposition required for such long-horizon launches is the first casualty of the degradation, consumed before capability peaks. If that ordering is wrong, if civilizations routinely retain, late into the degradation, both the capability and the residual coordination needed to seed even one self-sustaining artifact, then the silence does not follow, and the theory's central inference fails. This is the claim most in need of argument rather than assertion, and the place where the theory could most cleanly be shown to be wrong.

Fifth: evidence that the mechanisms are self-correcting at civilizational scale, that there are reliable endogenous processes by which the narrowing, the reflex, and the pacification reverse themselves before they reach critical levels. The theory claims the mechanisms are self-reinforcing and resistant to correction by the instrument itself. If this claim is wrong, if civilizations routinely develop compensating dynamics that interrupt the mechanisms before they become critical, the filter would not be as effective as the theory proposes.

Sixth: a demonstration that the quality of attention the theory identifies as the corrective capacity is not a real phenomenon, that what is described as a distinct mode of cognition outside the instrument's operation is actually just a variant of the same instrument, and that the distinction the theory is built on does not correspond to anything real. This would not merely weaken the theory. It would dissolve it. The theory is built on the reality of that distinction. If the distinction does not hold, neither does the theory.

These are genuine vulnerabilities. The theory holds them openly rather than explaining them away, because a theory that holds its vulnerabilities openly is a theory that can be improved, and improvement is the only honest relationship to have with a theory about something this important that cannot yet be confirmed.

IX. The Relation to Other Filter Candidates

The internal filter theory does not compete with external filter candidates in the way that two external candidates compete with each other. It occupies a different level of the analysis.

External catastrophe candidates, asteroid impact, nuclear exchange, engineered pathogen, ecological collapse, all operate at the level of events that happen to civilizations from outside or from specific developments within them. Each is contingent. Each depends on a specific chain of physical or social causation that may or may not obtain in any given case. They are real risks and should be taken seriously as such. They may be part of what filters some civilizations, in specific cases, under specific conditions. But none of them has the universality that the silence requires.

The internal filter operates at a different level: the level of the cognitive structure that makes civilization possible in the first place. It is not a risk that might or might not materialize. It is a structural consequence of the instrument exceeding its domain at scale, which is what the instrument does when it becomes powerful enough. It does not require a specific event. It does not require a specific mistake. It requires only the instrument continuing to do what the instrument does, at increasing scale, past the threshold at which the misapplication becomes civilizationally significant.

The self-replicating probe deserves specific treatment here, because it has already been used in this literature to draw the opposite conclusion. Tipler argued in 1980 that the absence of such probes implies not an internal filter but our solitude: a single prior civilization, anywhere, at any time in the galaxy's history, could have filled it with self-reproducing machines, and since we see none, none ever arose, and the filter lies behind us. The internal filter theory inverts the inference rather than denying the premise. The probes are absent, on this account, not because the builders never existed but because the window in which a degrading civilization could launch one is exactly what the filter consumes earliest. Tipler's argument and this one agree that the probe is the crux. They disagree about what its absence proves. That disagreement is precisely the exposed claim catalogued in Section VIII, and it is not resolved by being named here; it is only located.

The two kinds of candidates are not mutually exclusive. A civilization could face both an internal filter and external catastrophe risks. Indeed, the internal filter would tend to increase the likelihood of external catastrophe outcomes, because the progressive elimination of the quality of attention required to navigate wisely would reduce the civilization's capacity to recognize and respond to external threats. A civilization deep in the internal filter dynamic would be less likely to coordinate effectively against an engineered pandemic, less likely to sustain the collective

attention required for meaningful climate response, less likely to navigate nuclear risk through the kind of patient attention that its management actually requires. The internal filter does not replace external risks. It compounds them.

The theory also does not claim that the internal filter is the only possible internal filter. It claims that the specific mechanism it describes is a plausible candidate, that it has the universality that the silence requires if its two exposed premises hold, and that it is observably active in the one civilization we can examine. There may be other internal dynamics, arising from different features of technological civilization, that contribute to the filter. The theory is not exhaustive. It is one serious candidate, developed as fully and honestly as the available evidence permits.

X. The Present Moment

The theory has implications that are not merely theoretical for a civilization that is, on the theory's own account, currently undergoing the process it describes.

If the theory is correct, then the period we are in now, characterized by the mass deployment of artificial intelligence systems that amplify the mechanisms at unprecedented speed, by the measurable compression of the perceptual aperture at population scale, by the progressive substitution of representation for genuine engagement across every domain of collective life, is not a period of neutral technological development. It is a period in which the rate of the narrowing has sharply increased, in which the mechanisms are compounding faster than they have at any previous point in human history, and in which the conditions for the development of the corrective capacity are being consumed at an accelerating rate.

One implication of Section V is worth stating plainly rather than leaving implicit. The same systems now most aggressively amplifying the degradation are also the most plausible candidate for the persistent, substrate-independent automation that the theory's account of silence has to explain away. The technology that could, in principle, be the self-sustaining artifact that outlasts the attention required to launch it is, as actually built and deployed, the strongest accelerant of the attention collapse. Both descriptions are true at once, and which one weighs more depends on the ordering question Section VIII leaves open: whether the capacity for deliberate, long-horizon launches fails before or after raw capability peaks. The present moment does not answer that question. It only makes it concrete.

This is not an argument against technology. It is not an argument against AI. The theory does not claim the instrument is bad. It claims the misapplication of the instrument is structurally dangerous, and that the current moment represents a significant acceleration of that

misapplication. The instrument could, in principle, be oriented differently. The amplifier is not inherently pointed away from the quality of attention the theory identifies as the corrective capacity. It learned its current orientation from what it was trained on. That is a different problem than an amplifier inherently oriented that way, and the difference matters for what can be done about it.

What the theory implies is not despair and not a program. It implies a specific kind of attention to a specific kind of question. The question is not how to build better institutions or more sophisticated governance or more comprehensive analysis, though these things are not without value. The question is whether the quality of attention that the theory identifies as the corrective capacity can be developed and maintained in enough people in enough places before the conditions for it are fully consumed. That question is not answerable in advance. But it is the question that the theory, if taken seriously, puts at the center of the present moment.

The silence of the universe may be the answer to what happens when that question is not asked in time. Or it may be that we have not looked hard enough or long enough to hear what is out there. Both possibilities are real. The theory does not resolve them. It describes a mechanism, notes that the mechanism is running, and points toward the specific capacity whose development would determine, on this account, which silence the universe actually contains.

This theory is offered as a theory — a serious attempt to account for a real phenomenon using a coherent and evidentially grounded mechanism, with its premises stated, its vulnerabilities acknowledged, and its implications drawn out honestly. It is not offered as the truth. It is offered as something worth taking seriously, in a moment when the mechanism it describes is measurably accelerating, and in which the cost of not taking it seriously is, on the theory's own account, very high.

A.H.

April 2026

The companion investigation, *Two Kinds of Conditioning*, develops the mechanism in full detail and provides the evidentiary basis for the claims made here. This paper has compressed that development to focus on the theoretical application to the Fermi question. The two documents should be read together. Neither is complete without the other.

Note on Sources

The evidentiary basis for the four mechanisms, the biological conditioning framework, and the behavioral data cited in Section VI is developed fully in the companion investigation *Two Kinds of Conditioning*, which carries a complete bibliography. The Fermi paradox literature drawn on here includes Robin Hanson's original 1998 paper "The Great Filter: Are We Almost Past It?"; Nick Bostrom's "Where Are They? Why I Hope the Search for Extraterrestrial Life Finds Nothing" (MIT Technology Review, 2008); Stephen Webb's *If the Universe Is Teeming with Aliens, Where Is Everybody?* (2002, second edition 2015); and the broader literature on the Drake equation and SETI search parameters.

The argument in Section V regarding the cost of detectability and the self-replicating-probe objection draws on a further body of work. The energetic and observational cost of deliberate technosignatures is grounded in Freeman J. Dyson, "Search for Artificial Stellar Sources of Infra-Red Radiation" (Science, 1960), and Nikolai S. Kardashev, "Transmission of Information by Extraterrestrial Civilizations" (Soviet Astronomy, 1964). The self-replicating-probe argument and its use in the Fermi debate draw on John von Neumann's work on self-reproducing automata, completed and edited by Arthur W. Burks as *Theory of Self-Reproducing Automata* (1966); Ronald N. Bracewell, "Communications from Superior Galactic Communities" (Nature, 1960); and Frank J. Tipler, "Extraterrestrial Intelligent Beings Do Not Exist" (Quarterly Journal of the Royal Astronomical Society, 1980). The counter-argument that an advanced civilization would refrain from saturating the galaxy with such probes is Carl Sagan and William I. Newman, "The Solipsist Approach to Extraterrestrial Intelligence" (Quarterly Journal of the Royal Astronomical Society, 1983), repurposed here to support the claim that the disposition required for long-horizon launches is consumed early by the degradation.

The distinction between internal and external filter candidates is implicit in much of this literature but has not, to this author's knowledge, been developed in the specific direction this paper pursues. The treatment of the self-replicating probe as an internal-filter problem rather than, as in Tipler, evidence of our solitude, is original to this investigation, as is the mechanism proposed here.